2장. 통계적 추론이란?

◈ Introduction

부동산 예측 전문가가 "작년 1월엔 집값이 50% 올랐고, 2월엔 15% 내렸고, … , 12월엔 10% 내렸으니 올해도 이에 맞게 투자하면 됩니다." 라고 말하면 과연 신빙성이 있을까? 올해 가격 추세가 작년의 추세와 동일하게 움직인다는 보장은 거의 없다. 이처럼 구구절절하게 말하는 것 보다는 투자자에게 "작년은 평균적으로 10% 올랐으니, 올해도 비슷하게 오를 겁니다." 라고 말하는 전문가가 더 신빙성이 있어 보일 것이다.

모델을 통한 예측도 앞선 경우와 같다. 과거 데이터의 모든 관측치들에 세세하게 신경을 쓰면 예측은 틀릴 가능성이 높다. 또한 이상점(Outliers)들이 데이터에 포함되어 있으면 예측을 방해할 수도 있다. 좋은 예측 모델은 간단 명료해야 한다. 그런 모델들은 현상들을 쉽게 설명하고 좋은 예측력을 갖는다. 우리가 이번에 배울 Test MSE를 통해 예측력이 좋은 모델이 무엇인지 알려줄 것이다.

1.Systematic Information

X 를 입력과 관련된 독립변수, Y 를 출력과 관련된 종속변수, 그리고 ε 를 오차항이라고 할 때 모델은 다음과 같이 표현된다.

$$Y = \underbrace{f(X)}_{\text{Systematic Information}} + \underbrace{\varepsilon}_{\text{Random Error}}$$

$f(X)$를 systematic information이라고 한다. 즉, $f(X)$는 종속변수와 독립변수 사이의 관계를 나타내는 체계적인 정보이다. $f(X)$를 추정한 $\hat{f}(X)$을 구하게 되면 우리는 비로소 추론 혹은 예측을 할 수 있다. 모델이 선형회귀모형을 통해 표현되는 경우, 독립변수의 계수들을 통해 우리는 종속변수와 독립변수의 관계를 추론할 수 있고, $\hat{f}(X)$ 에 주어지지 않은 X 의 정보들을 대입하면 이에 대한 값 Y 를 예측할 수 있다. 그렇다면 우리는 어떻게 $\hat{f}(X)$을 얻을 수가 있을까? $\hat{f}(X)$ 를 통해 구해진 예측값 $\hat{Y}$ 이 Y 와 최대한 비슷하면 된다. 다르게 말하면, 모델을 통한 예측값과 실제값의 오차가 최소가 되어야 한다. 하지만 오차를 완전히 없애는 것은 불가능하다. 오차에는 줄일 수 없는 부분이 있기 때문이다.

2. Reducible Error, Irreducible Error

오차는 reducible error와 irreducible error로 나뉜다. Reducible error는 줄일 수 있는 오차이다. 추정한 $\hat{f}(X)$가 실제 $f(X)$에 가까울수록 추정에 의한 오차가 감소하기에 더 정교한 추정방법을 사용하면 reducible error는 감소한다. 이와 달리 irreducible error는 감소시킬 수가 없다. 그 이유로는 첫째, 종속변수를 설명하는데 도움이 되는 독립변수의 부재이다. 설명력이 있는 독립변수가 $f(X)$ 에 포함되지 않기 때문에 $f(X)$ 의 좋은 추정으로는 해결이 불가능한 오차가 발생한다. 둘째, 측정대상의 무작위성이다. 측정 대상이 사람일 경우, 측정 대상의 기분, 태도, 신체적 정신적 상태에 따라 측정값이 변동될 수가 있다. Irreducible error와 reducible error는 다음과 같이 수리적으로도 표현 가능하다.

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\varepsilon-\hat{f}(X)]^2 \\
&= [f(X)-\hat{f}(X)]^2 + Var(\varepsilon) \\
&= reducible + irreducible
\end{aligned}
$$

3. Parametric Methods vs Non-parametric Methods

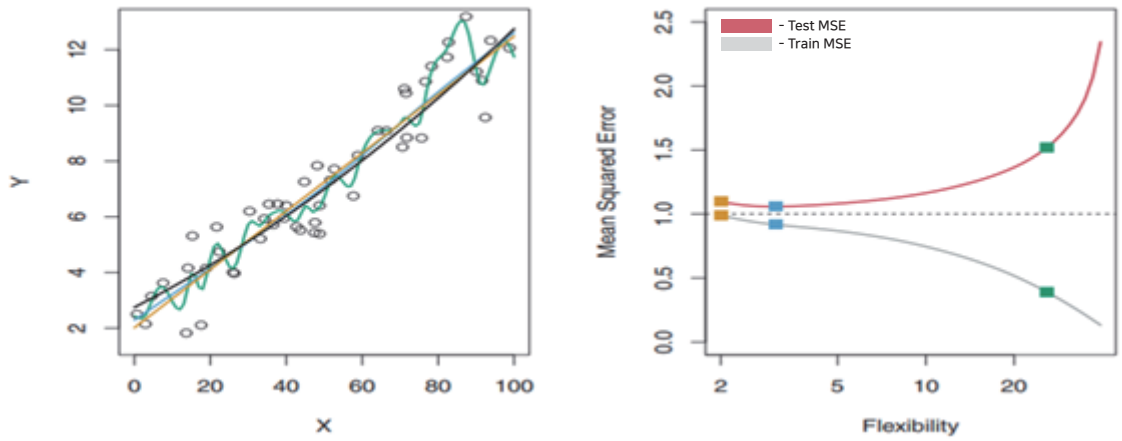
$\hat{f}(X)$를 구하는 방법은 크게 Parametric Methods과 Non-Parametric Methods으로 나뉜다. Parametric Methods는 $f(X)$ 에 함수 또는 모양을 사전에 가정하고 Non-Parametric Methods에서는 가정하지 않는다. Parametric 방법의 예시로 선형회귀모형을 생각해보자. 선형회귀모형에서 $f(X)$의 형태가 독립변수들로 이루어진 선형모델일 것이라고 가정을 하고 각 독립변수의 계수를 추정하는 것을 목적으로 한다. 우리가 사전에 $f(X)$의 형태를 가정했기에 선형회귀모형은 Parametric Method 이고 이로 인한 장단점이 생긴다. 장점으로는 사전에 $f(X)$ 의 형태를 가정했기에 $f(X)$ 자체를 추정하지 않고 독립변수들의 계수만 추정하면 되기에 Non-Parametric Method 보다 손쉽게 사용 가능하다. 하지만 인위적으로 $f(X)$의 형태를 가정하기에 만일 실제 형태가 우리가 가정한 형태와 많이 다르면 계수 추정과 상관없이 좋지 않은 모형이 만들어진다. 이를 방지하기 위해서 유동성이 좋은 함수 또는 모형을 가정한다.

Non-parametric Methods의 예시로는 Thin-plate Spline이 있다. Thin-plate Spline 방법은 부드러움의 정도(smoothness)만을 인위적으로 통제하고 Parametric model와는 다르게 모양에 대한 가정이 없다. 부드러움의 정도를 작게 설정하면 데이터에 더 적합한 모델이 나오게 된다. Non-parametric Methods은 $f(X)$에 함수나 모양에 가정을 하지 않아도 되는 장점이 있지만 $f(X)$의 함수자체를 추정해야 하기 때문에 Parametric Methods 보다 많은 관측치 데이터를 필요로 하는 단점이 있다.

4. Training MSE vs Test MSE

Parametric Methods와 Non-Parametric Methods에 대하여 설명한 부분을 읽다가 의문이든 사람들이 있을 것이다. Parametric Methods에서는 유동성이 높은 방법을 사용하면 실제 $f(X)$와 차이를 감소시킬 수 있다고 언급하였다. 그리고 Non-Parametric Methods의 예시에서 부드러움의 정도를 작게 하면 데이터에 더 적합한 모델이 나온다고 하였다. 그렇다면 데이터에 잘 적합하게 가장 유동적인 모델을 쓰거나 부드러움의 정도를 많이 작게 하는 방법만을 사용하면 되지 않겠는가?

우리가 모델을 만드는 주목적은 우리가 만든 모델이 데이터와 얼마나 잘 적합한지가 아니라 얼마나 추정을 잘하는지에 있다. 추정한 모델과 모델에 사용된 학습데이터(Training data)의 오차가 작을수록 추정된 모델과 추정에 사용되지 않은 평가데이터(Test data)의 오차가 작아진다고 생각할 수도 있다. 하지만 항상 그런 것은 아니며 특정 기점부터는 모델과 학습데이터의 오차가 작아질수록 모델과 평가데이터의 오차가 커진다. 이 현상을 과적합(overfitting)이라고 하며 모델학습 과정에서 무작위로 발생하는 오차항들까지도 적합(fitting)을 하여 이런 현상이 발생한다. 학습데이터에서 실제값과 모델을 통한 추정값의 오차의 제곱합을 Training MSE라고 하고 평가데이터에서 실제값과 모델을 통한 추정값의 오차의 제곱합을 Test MSE라고 하면 [그림1] 같은 관계가 성립한다.



[그림1] 유동성(Flexibility)에 따른 Fitting의 변화 (좌측) 및 Train/Test MSE의 변화 (우측)

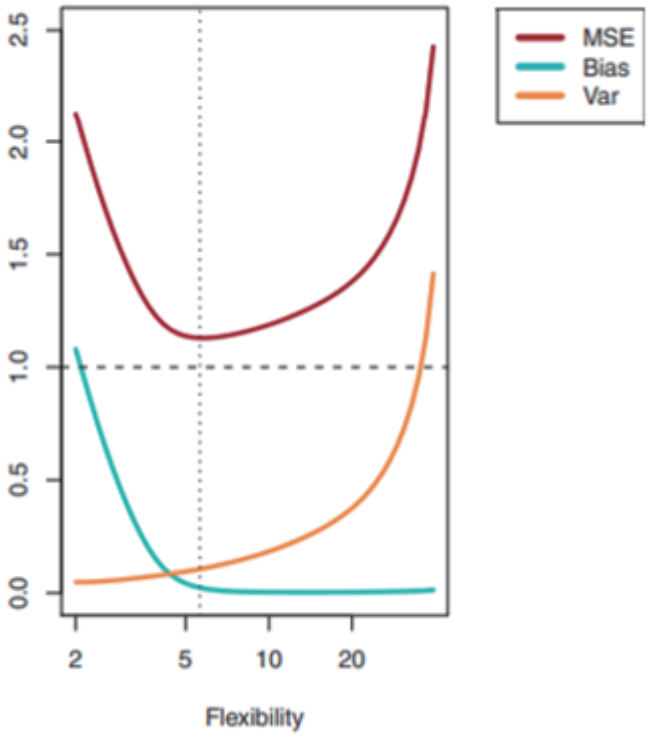
[그림1] 좌측에서 모델이 유동적일수록 데이터에 더 적합한 것을 확인할 수 있다. (초록색 선은 곡선형태로 많은 데이터들을 fitting 하고 있다.) 또한 우측 그림에서 모델이 유동적일수록 Training MSE는 계속 낮아지는 반면 Test MSE는 낮아지다가 어느 순간부터는 점점 커지는 과적합 현상을 수치적으로 확인할 수 있다. Test MSE가 낮아지다가 커지는 U 타입의 그래프를 나타내는 이유는 Bias-Variance trade-off로 설명 가능하고 [그림2]에서 확인 할 수 있다.

5. Bias-Variance Trade-Off

Test MSE는 다음과 같이 분해 가능하다.

$$E(Y-\hat{f}(X))^2 = Var(\hat{f}(X)) + [Bias(\hat{f}(X))]^2 + Var(\varepsilon)$$

즉, Test MSE는 추정치의 분산, Bias 그리고 Irreducible error의 합이다. 모델의 유동성이 클수록 추정치의 bias는 커지고 분산은 작아지고 유동이 작아질수록 추정치의 bias는 작아지고 분산은 커지며 이를 Bias-Variance trade-off이라고 한다. 즉, 모델의 유동성에 따른 bias의 변동성과 분산의 변동성의 크기 정도에 따라 Test MSE가 작아지다가 커지는 현상이 발생한다. [그림 2]에서 모델의 유동성에 따라 초반에는 bias가 분산이 커지는 것에 비해 상대적으로 많이 감소하여 Test MSE가 감소하는 것을 확인할 수 있다. 또한 모델의 유동성이 커지면서 분산이 bias에 비해 상대적으로 많이 증가하여 결국 Test MSE가 다시 증가하는 U 모양의 그래프로 표현된다.



[그림2] 유동성(Flexibility)에 따른 Test MSE의 변화 (Bias-Variance Trade-off을 중심으로)

★ 쉬어가는 페이지 ★

머신러닝의 골칫거리: 과적합(overfitting)

과적합(overfitting)이란 통계나 머신러닝에서 사용되는 용어로, 제한된 샘플이나 학습 데이터에 지나치게 특화되어 새로운 샘플에 대한 예측 결과가 오히려 나빠지거나 학습의 효과가 나타나지 않는 경우를 말하며, 일반화를 떨어뜨리는 주범 중 하나이다. 학습 데이터는 충분히 잘 설명할 수 있지만, 새로운 데이터에 대한 해석을 제대로 할 수 없게 되어, 머신러닝을 무용지물로 만들어 버린다. 반면에 과소적합(underfitting)은 지나친 단순화로 인해 데이터에서 뚜렷한 특징을 제대로 찾지 못하여 새로운 데이터는 고사하고 학습 데이터에 대한 해석조차도 엉터리가 되는 경우를 말한다.

예를 들어, 개를 잘 인식할 수 있는 머신러닝 모델을 개발하고자 한다면, 지구상에 존재하는 다양한 모양의 개를 학습 데이터로 준비를 해야 한다. 예를 들어, 진돗개, 쉐퍼드, 치와와 종만을 학습 대상으로 삼는다면, 달마시안이나 불독 종은 개라고 인식하지 못할 수도 있는데 이런 상황을 overfitting이라고 부른다. 또한, 몸에 털이 있고 4개의 다리를 가진 짐승으로 지나친 단순화로 한다면, 고양이도 개라고 인식하는 경우를 underfitting이라고 부른다.

과적합 문제가 발생하는 원인은 여러 가지이다. 기본적으로는 학습 데이터가 너무 치우쳐 있거나 데이터의 양이 부족하여 전체를 설명하기가 어려운 경우이다. 이런 경우는 학습 데이터 양을 늘리는 것이 해결책일 수 있지만, 학습 데이터를 늘리는 것은 막대한 비용과 노력을 필요로 한다. 또한 데이터에 대한 정확한 이해가 없다면, 표본이 치우쳤는지 아니면 전체를 잘 대변하는지 파악하기가 힘들기 때문에, 보통은 학습용 데이터 집합(training data set)과 검증용 데이터 집합(validation data set)을 나누고, 학습에 사용하지 않은 검증용 데이터 집합을 이용하여 학습이 잘 진행되었는지를 판단한다.

★ 쉬어가는 페이지 ★

또 다른 원인으로는 너무나 복잡한 모델로 설명하고자 하는 경우이다. 예를 들어, 약간의 잡음(noise)을 포함하고 있지만 '2차 함수'로 대표하면 충분히 설명 가능한 경우를, 각 개별 점들에 너무 집중하여 'N차의 고차 함수(여기서 N은 샘플의 개수)' 표현을 한다면, 학습용 데이터에 대하여 100% 정확한 해석을 할 수 있겠지만, 새로운 데이터에 대하여 큰 오류를 보일 가능성이 있다. 실제로 학습용 데이터는 데이터를 만드는 시스템이나 사람의 실수로 인하여 잡음이나 오류가 발생할 가능성이 늘 있는데 너무 학습 데이터에만 특화 된다면 전혀 엉뚱한 모델이 나올 수 있다. 이런 경우에 항상 과적합과 일반화의 균형(trade-off) 이슈가 생길 수 있으며, 보통은 '오캄(Occam)의 면도날' 방법을 적용하여 단순화 시키는 것을 권장하기도 한다. 오캄의 면도날은 14세기 영국의 신학자이자 논리학자인 오캄의 저서에 등장하며, 중세의 철학자와 신학자들이 복잡한 논쟁을 배제시키자는 뜻에서, 설명이 더 복잡한 이론은 배제할 때 흔히 사용이 된다. 단순한 모델을 선택하기 위한 방법으로론 보통 '정규화(regularization)' 방법이 사용이 된다.

출처 : YTIMES 와이타임스(http://www.ytimes.co.kr)
[이석중의 알기쉬운 인공지능]